Machine Learning Engineer Nanodegree
Capstone Proposal

Sonja Tilly

April 14, 2017

Proposal

Domain Background

Lending Club is the world's largest peer-to-peer online platform, which has enjoyed tremendous success since its foundation in 2007.

Borrowers are able to apply for loans, which are rated – the lower the rating, the higher the interest rate. Investors can purchase notes of those loans. As borrowers make payments on their loans, investors are paid monthly on their respective notes, less the impact of any defaults and Lending Club fees. According to Lending Club, the estimated annual net return on a diversified loans portfolio is 5%[i], which remains an attractive return in the current low interest rate environment, with the 5Y US Treasury yield below 2%.

A number of research pieces have been published by various bodies, each examining different aspects of the dataset.

Research published on the NY Data Science Academy website segments the Lending Club dataset 'good' and 'bad' loans and applies a logistic regression model to estimate the expected loss on loans listed as 'current'[ii]. BlackMoon Financial Group analyzed non-numerical data and demonstrated how this information can help investors make better decisions when manually selecting loans for their portfolio[iii]. Researchers from Stanford University used supervised learning techniques to predict the probability that a loan application will get approved. Further, they used various visualization techniques to show patterns in the dataset[iv]. A paper by a student at General Assembly determined the key features indicative of loan defaults and built models to predict them[v].

Working in Asset Management with focus on fixed income strategies, I found this dataset particularly interesting as it may hold deeper insights into the dynamics of a relatively new yet fast-growing market segment and how investors could enhance their asset allocation when manually constructing a portfolio of loans and ultimately achieve a better return on their investment.

Problem Statement

Lending Club allows investors to purchase notes in loans based on their credit rating. However, investors have no visibility as to the probability of default of any one loan. Having this information could significantly enhance investors' asset allocation strategy as they would be able to select those loans with a lower probability of default given the same interest rate level.

Datasets and Inputs

The dataset obtained from Kaggle[vi] contains complete loan data for all loans issued between 2007 and 2015, including the loan status (current, late, fully paid, etc.) and payment information. The data contains a range of different data types such as numerical data, free form text and date objects, which means that the raw data will require preprocessing. The file is a table of 887,379 observations and 74 variables. A data dictionary is provided in a separate file.

The dataset will be used to examine the relationships between loan performance and other variables.

Solution Statement

In this project, I will visually explore the dataset to gain deeper insights into the relationships of different variables with loan performance. I intend to investigate questions such as:

- How has the amount and quality of loans issued evolved over time?
- How has loan performance varied over time?
- What is the relationship between interest rate and loan performance?
- What is the relationship between loan grade and loan performance?
- What are the main loan purposes? How are they related to loan performance?
- How is occupation related to loan performance?
- How is home ownership related to loan performance?
- How is loan performance related to applicant state?

Further, I will apply supervised learning methods to predict the probability of a bad loan (i.e. charged off, default or late) and determine the method, which will produce the best result as defined by the AUC score.

Benchmark Model

The benchmark for this project is a simple logistic regression model, appropriate as the project represents a classification problem. The model is used to estimate the probability of a bad loan. Benchmark model performance is measured by the AUC score.

Evaluation Metrics

The dataset is unbalanced, with bad loans accounting for only a small portion of the total number of loans. Due to this class imbalance, model accuracy will be assessed the Area Under the Curve (AUC).

A receiver operating characteristic curve or ROC curve is a graph plotting the true positive rate against the false positive rate at different thresholds. It indicates how well a classifier can discriminate positive and negative instances and identify the best threshold for discriminating them. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

- TPR (also called Recall or Sensitivity) = sum total positives/sum of condition positive
- FPR (also called Fall-out) = sum of false positives/sum of condition negative

The area under the curve (AUC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one[vii].

Project Design

Programming language and libraries

- Python, Sci-kit learn, TensorFlow

Data preprocessing

The dataset will be uploaded into a Python environment. Since it is a very large dataset, a smaller subset of the data may have to be used.
The Lending Club dataset contains various data types such as numerical data, date objects and text. Some variables such as loan purpose or job title are in text format and have to be converted. Further preprocessing required for modelling could include scaling data, transforming categorical variables and dealing with null values.

Data visualization

At this stage, the relationships between different variables and loan performance will be explored visually. Preliminary findings will be discussed and summarized.

Feature selection

Insights from the above may be helpful when selecting features that are relevant to predicting loan performance.

Modelling

Build benchmark model. A logistic regression model will be built and model performance will be assessed using the AUC score.

Different models will be constructed using algorithms such as Naïve Bayes, RandomForests and Neural Networks. All three algorithms are suited to a classification problem. The dataset includes a large number of observations, particularly required by the RandomForest and Neural Network algorithms. Model performance will be assessed using the AUC score compared to that of the benchmark model.

Model optimization methods such as GridSearch will be applied to improve performance. Optimized model performance (including benchmark model) will be compared to the original models and the benchmark model.

Conclusion

Modelling results will be discussed and the most appropriate model will be identified. Discussion will reference the results to the problem statement and their potential impact on asset allocation strategies.

[i] https://www.lendingclub.com/investing/alternative-asset-investments/how-it-works
[ii] http://blog.nycdatascience.com/r/p2p-loan-data-analysis-using-lending-club-data
[iii] https://blackmoonfg.com/site/pdf/bm_non_numerical_data.pdf
[iv] http://Pcs229.stanford.edu/proj2016spr/report/039.pdf
[v] https://res.cloudinary.com/general-assembly-profiles/image/upload/v1416535475/uwumooopplttsmpgu1goo.pdf
[vi] https://www.kaggle.com/wendykan/lending-club-loan-data
[vii] www.wikipedia.com